Products and Services

IDC DOCUMENTATION

Formats and Protocols for Continuous Data

CD-1.0

Approved for public release;
distribution unlimited

Revision 0.1

Notice

This document was published by the Monitoring Systems Operation of Science Applications International Corporation (SAIC) as part of the International Data Centre (IDC) Documentation. It was first published in May 1998 and was republished electronically as Revision 0.1 in February 2002 to include minor changes (see the following Change Page). IDC documents that have been changed substantially are indicated by a whole revision number (for example, Revision 1).

Contributors

Richard Stead, Science Applications International Corporation
Jerry Carter, Science Applications International Corporation
Warren Fox, Science Applications International Corporation

Trademarks

Ordering Information

The ordering number for this document is SAIC-01/3054.

This document is cited within other IDC documents as [IDC3.4.2Rev01].

Change Page

This document is Revision 0.1 of the Formats and Protocols for Continuous Data CD-1.0. The following changes have been made for this publication:

Page	Change
Cover Page	Added “CD-1” to title to compare with [IDC3.4.3Rev0.1].
as needed	Added change bars (vertical lines that appear in the margin) to identify new or revised material.
as needed	Minor revisions to wording to better explain, remove ambiguity, and fully represent the standard as implemented.
as needed	Changed “sender” and “receiver” to data producer (dp) and data consumer (dc), consistent with usage in other IDC documentation.
as needed	Clarified the usage of “address” and “port”.
Notice Page	Changed notice page to include an explanation of the Revision number, a reference to the Change Page, new trademarks, and a new SAIC number.
Change Page	Added this change page.
ii	Added a description of changes to the document.
iv	Eliminated a reference to the Roadmap under “Using this Document”.
vi	Added/deleted several technical terms in Table III.
2	Expanded discussion of time-outs.
3	Added description and table of time-outs.
8	Added introductory paragraph.
8	Clarified discussion of IEEE representation, byte ordering and byte alignment (also modified elsewhere as needed).
8	Added section on standards for number and time representation.
10	Added footnote describing earlier format frame.
11	Added footnote on unused channels.
11	Added footnote on calibration units.

Page	Change
12	Altered frame length discussion to be consistent with stated usage in other documents.
13	Altered discussion of description field and added a footnote, partly addressing weather data.
14	Fixed incorrect statement on subframe time stamp.
17	Added discussion of delay intervals between retries.
18	Altered example to be more typical (10 seconds), corrected minor deviations from the standard.
20	Expanded discussion of example data frame.
24	Added statement regarding missing samples.
29	Added chapter on Policies.
36	Added sentence on overlapping data.
37	Added references for time, floating-point, and compression standards.
G1–G4	Added definitions for "array", "byte alignment", "calibration", "Canadian compression", "client", "connection", "continuous waveform data", "count(s)", "data consumer", "data provider", "element", "epoch time", "FDSN", "frame", "hydroacoustic", "infrasonic (infrasound)", "IP address", "ISC", "LIFO", "NEIC", "nil", "Pa", "port", "sample", "seismic", "sensor", "server", "site", "station", "time series", and "UNIX".

Formats and Protocols for Continuous Data CD-1.0

CONTENTS

Formats and Protocols for Continuous Data CD-1.0

FIGURES

Formats and Protocols for Continuous Data CD-1.0

TABLES

About this Document

This chapter describes the organization and content of the document and includes the following topics:

- Purpose
- Scope
- Audience
- Related Information
- Using this Document

About this Document

PURPOSE

This document describes the standard International Data Centre (IDC) formats and protocols for International Monitoring System (IMS) continuous data exchange.

This document is Revision 0.1 of Formats and Protocols for Continuous Data CD-1, originally published May, 1998. The following changes have been made for this revision:

- General

  The publication date and document number in the footers for all main pages of the document were changed to reflect the revision. Change bars (vertical lines that appear in the margin) were added to identify new or revised material. "Sender" and "receiver" were changed to data producer (dp) and data consumer (dc) throughout the document to be consistent with usage in other IDC documentation.

- Front material

  The Notice Page was changed to include an explanation of the revision number, a reference to the Change Page, and a new SAIC document number. A Change Page, which summarizes technical revisions to the document, was added after the Notice Page.

- About this Document

  A description of changes to the document was added, the reference to the Roadmap under "Using this Document" was eliminated, the Conventions were updated by eliminating many of the standard conventions not used in this document, and the Technical terms were updated.

- Chapter 1: Continuous Data Protocol

  A thorough discussion of "time-outs" was added.

- Chapter 2: Frame Formats

  The Introduction for this chapter was updated and a new section on standards was added. Many changes to the text were made, primarily as footnotes, to clarify the use of early Data Format Frames, unused channels, calibration units, overlapping data, frame lengths, description fields (and their use for weather data), uncompressed data formats, and delay intervals between retries.

- Chapter 3: Data Formats

  A discussion of delay intervals between retries was added, the discussion of the example data frame was expanded, and a statement regarding missing samples was added.

- Chapter 4: Policies for Continuous Data Transmission

  A "Policies" chapter was added to clarify the use of the protocol.

- References

  New references were added to support the revisions made in the document.

- Glossary

  Several new terms were added to the Glossary.

- Index

  Several new index terms were added. New index terms do not have change bars.

SCOPE

This document includes a description of the continuous data protocol, frame formats, and data formats. The real time data exchange model includes several formats that are used to describe the data being sent, communicate changes, and forward the data. This document does not describe the software for receiving or forwarding continuous IMS data. These topics are described in sources cited in "Related Information."

AUDIENCE

This document is intended for software developers, engineers, and users of the continuous IMS data exchange systems.

RELATED INFORMATION

The *Formats and Protocols for Continuous Data CD-1.1, Revision 0.2* [IDC3.4.3Rev0.2] describes alternative formats and protocols for continuous data.

See "References" on page 37 for a listing of all the sources of information consulted in preparing this document.

USING THIS DOCUMENT

This document is part of the Products and Services category of the overall documentation for the IDC. The Products and Services category provides descriptions of IDC products and their formats.

This document is organized as follows:

- Chapter 1: Continuous Data Protocol

  This chapter provides a high-level description of the protocol that is followed to establish a connection and transmit continuous data.

- Chapter 2: Frame Formats

  This chapter describes the formats of the frames that are used to establish connections, transfer data, and alter connections.

- Chapter 3: Data Formats

  This chapter provides the formats of the data that are transferred.

- Chapter 4: Policies for Continuous Data Transmission

  This chapter describes the policies that govern the correct implementation of the protocol.

- References

  This section lists the sources cited in this document.

- Glossary

  This section defines the terms, abbreviations, and acronyms used in this document.

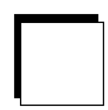
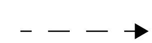
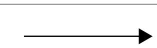
- Index

  This section lists topics and features provided in this document along with page numbers for reference.

Conventions

This document uses a variety of conventions, which are described in the following tables. Table I shows the conventions for data flow diagrams. Table II lists typographical conventions. Table III explains certain technical terms that are not part of the standard Glossary, which is located at the end of this document.

TABLE I: DATA FLOW SYMBOLS

Description	Symbol[1]
external source or sink of data	
control flow	
data flow	

1. Most symbols in this table are based on Gane-Sarson conventions [Gan79].

TABLE II: TYPOGRAPHICAL CONVENTIONS

Element	Font	Example
titles of documents	*italics*	*Operational Manual for the International Data Centre*
computer code and output	`courier`	`ts.idc.org:80000`

TABLE III: TECHNICAL TERMS

Term	Description
compressed data	data that have been reduced significantly in size, without loss of information, to make transmission more efficient
format	data structure used by application programs to pass information across an interface
forward byte ordering	byte order for integers and floats in which the least significant bit is the right-most bit, also referred to as "big endian" in other computer literature
hexadecimal	base 16 numbering system; A = 10, B = 11, and so on
`inetd`	UNIX process that waits for network connections
network address	either the IP address or the fully-qualified host name of the computer at the IP address (for example, "`ts.idc.org`"), which can be looked up to find the IP address
port or port number	16-bit number indicating an address for one TCP (or UDP) queue on a computer; each TCP/IP virtual circuit is defined by the IP address of each of the two computers and the port used on each computer for that one circuit
protocol	set of conventions that establishes the sequence of frames transmitted between the data provider (dp) and the data consumer (dc)
time-out	a pre-defined length of time that an application will attempt an action or wait for an event, after which the application will terminate the action and re-initialize the process

Chapter 1: Continuous Data Protocol

This chapter provides an overview of the Continuous Data protocol (CD-1 protocol) used to establish a connection, continuously transmit data, and alter a connection. The following topics are included:

- Introduction
- Establishing Connections
- Transmitting Data
- Altering Connections
- Terminating Connections

Chapter 1: Continuous Data Protocol

INTRODUCTION

The CD-1 protocol for sending and receiving continuous data in near-real time provides a reliable method for transmitting and receiving digital time series data. Up to 100 data channels from a station or array of stations can be transmitted using a single connection. The data may be provided in compressed or uncompressed formats and with or without authentication signatures (see "Formats and Protocols" in [GSE95b]).

The protocol uses units of information called "frames" to establish or alter a connection and to exchange data. Each frame conveys a specific type of information between the data producer "dp" and the data consumer "dc." The protocol is the policy for exchanging frames. This policy must be followed precisely or the connection is terminated automatically by one of the parties.

Figure 1 indicates the succession of frames in a normal data exchange sequence. Time in this figure flows from top to bottom. At any instance, one frame at most is being transmitted or processed. The first two frames (connection and port assignment) are used to establish a connection. The remaining frames effect the data transfer.

As a safeguard against lost communication and similar failures, the sending and receiving software uses a "time-out" to determine if a connection is lost. If frames cannot be sent or are not received within the time-out periods established by the dp and the dc, the connection is assumed to be lost and a new connection must be established by the dp. The time-outs to be enforced by both the dp and dc are given in table 1. Note that these time-outs are enforced on application-level activity. Underlying activity on the transport, network or link layers is not considered.

Any changes to these time-outs, if required by special cases, would need to be negotiated in advance between the dc and dp, and are discouraged. Dcs and dps are not assumed to allow these values to be altered.

TABLE 1: TIME-OUTS

frame type or action	time-out (seconds)
Connection Request Frame	90
Port Assignment Frame	60
Data Format Frame	600
Data Frame	600
Alert Frame	600
connecting action (initial)	60
re-connecting action to designated port	600
terminating action	0

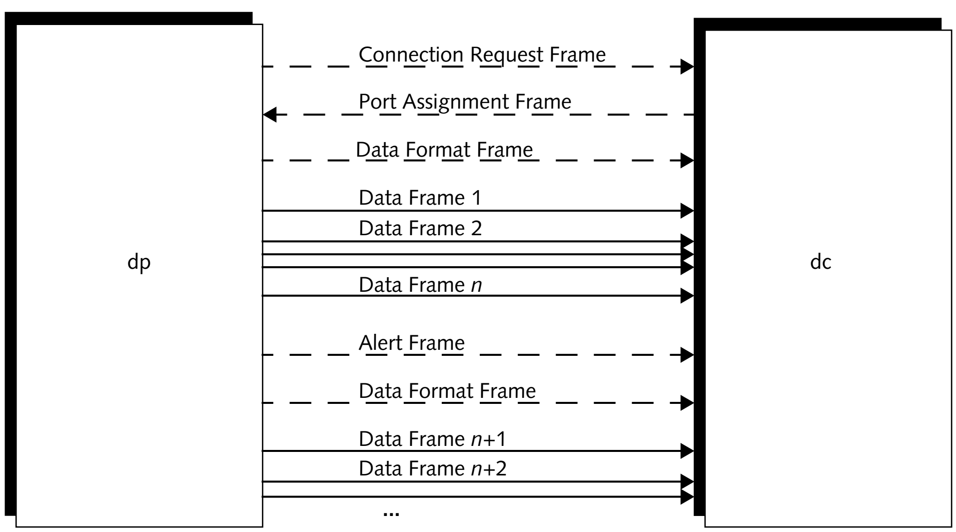


FIGURE 1. CONTINUOUS DATA COMMUNICATIONS PROTOCOL CD-1

ESTABLISHING CONNECTIONS

The facility sending data (data producer, dp) is responsible for initiating the connection with the facility receiving data (data consumer, dc). (In client/server terminology, the dp is the client and the dc is the server.) To initiate communication, the dp attempts to open a connection to the dc on a predesignated network address and port (for example, `ts.idc.org:8000`). The dp announces itself to the dc by sending a Connection Request Frame that contains a station code. The format of this frame and other frames are described in detail in "Chapter 2: Frame Formats" on page 7.

A program at the dc, such as the UNIX process `inetd`, must be running and listening to the designated port at all times. When a Connection Request Frame is received, the dc verifies the source of the dp request and provides a new port and IP address in a Port Assignment Frame. When the dp receives the IP address and port assignments, it drops the original connection to the dc and connects to the assigned IP address and port. This connection is subsequently used for all data transfers.

TRANSMITTING DATA

After the dp has connected to the dc at the IP address and port assigned by the dc, the dp sends a Data Format Frame, which describes the format of the subsequent Data Frames. The dp can then send Data Frames until the data no longer match the previous Data Format Frame. This procedure, in which transmission is initiated by the dp, is known as the "push" model. The dc does not poll the data. The normal TCP/IP error correction and packet ordering apply.

ALTERING CONNECTIONS

Connections may be altered through the exchange of Alert Frames. These frames may be sent by either the dp or the dc. The dc uses Alert Frames to notify the dp to use a different IP address and port or to cease sending data. The dp uses Alert Frames to notify the dc that the communication will cease or to expect a new Data Format Frame.

TERMINATING CONNECTIONS

Connections may be intentionally terminated by either the dp or dc. The party terminating the connection sends an Alert Frame, and then drops the connection, as described above.

Connections may be unintentionally terminated due to a slow or failed dp, dc, or communications system. Unintentional terminations are detected by either the dp or the dc after a period known as the “time-out” (see Table 1). The surviving party lacks the information to determine the cause of a time-out, but reacts the same in all cases by also dropping the connection, if it has not already done so via its own time-out detection.

Typically, after a connection is established, it remains active and in use for days, weeks, or months until either a failure (dp, dc, or communications system) or the dp or dc terminates the connection for maintenance or reconfiguration.

Chapter 2: Frame Formats

This chapter describes the formats for continuous data exchange and includes the following topics:

- Introduction
- Connection Request Frame
- Port Assignment Frame
- Data Format Frame
- Data Frame
- Alert Frame
- Example

Chapter 2: Frame Formats

INTRODUCTION

A frame of the protocol is an interface data structure. As such it is defined by its format, which explains the meaning of the sequence of bytes passed between the sending application and the receiving application.

The connection and transmission protocol CD-1 uses five distinct types of frames: Connection Request Frame, Port Assignment Frame, Data Format Frame, Data Frame, and Alert Frame. The formats for these frames are described below.

STANDARDS

The CD-1 protocol adheres to standards for number and time representation as described in the following paragraphs.

Representation of Numbers

The CD-1 format uses the Institute for Electrical and Electronic Engineers (IEEE) standard for numerical representation. IEEE integers (or shorts or longs) having four (or two or eight) bytes are used exclusively for fixed point numbers. IEEE floats are used for floating point numbers. The IEEE representation implies forward byte ordering of multi-byte values (also known as “network byte order”) and is used natively on Motorola and Sun computers, but not Intel Central Processing Units (CPUs). Values generated on machines that do not use IEEE as their native numerics may need to swap bytes before transmitting data.

Representation of Strings

Strings are represented with the ASCII character set and are padded as described in the following sections. These strings are only NULL-terminated if so dictated by the padding rules.

Byte Alignment

All portions of this format that are of variable length have a length that is a multiple of four. A variable-length field that does not have a length that is a multiple of four must be padded to the next multiple of four with trailing bytes set to NULL (0).

Epoch Time

The time used by the CD-1 protocol is epoch time, and is represented as an 8-byte IEEE floating point value indicating the number of seconds since the epoch (1 January, 1970 00:00:00), not including leap seconds.

CONNECTION REQUEST FRAME

To initiate transmission of data, the dp forwards a Connection Request Frame to the dc. Table 2 defines the Connection Request Frame.

TABLE 2: CONNECTION REQUEST FRAME

Field	Format	Description
station name	8-byte string	registered station code

PORT ASSIGNMENT FRAME

To assign a port for data transmission, the dc sends a Port Assignment Frame to the dp. Table 3 defines the Port Assignment Frame.

TABLE 3: PORT ASSIGNMENT FRAME

Field	Format	Description
IP address and port	32-byte string	IP address number and port number separated by "/"

DATA FORMAT FRAME

The Data Format Frame describes the format of the time series data that are to be transmitted. The dp forwards a Data Format Frame whenever communications are first established or re-established following a disconnection. A Data Format Frame is also sent whenever the format of the data is changed (for example, after a change in calibration or when a channel is dropped). The Data Format Frame is 2,820 bytes long[1]. The first 20 bytes provide information about the Data Format Frame itself and general information about the Data Frames that will follow. The last 2,800 bytes specify channel information for up to 100 channels. The order in which the channels appear in the Data Format Frame is the same order that must be followed for the data in the Data Frames that follow. Table 4 specifies the Data Format Frame.

TABLE 4: DATA FORMAT FRAME

Field	Format	Description
frame identification	4-byte integer	identification for Data Format Frame = 2
frame length	4-byte integer	length in bytes of Data Format Frame
maximum frame size	4-byte integer	maximum size of following Data Frames in bytes (max ≤ 10,000,000; divisible by 4)

1. Early versions of *libalpha* that implemented CD-1 did not specify calibration and period in the Data Format Frame. A dc operating with legacy dps may be required to accept Data Format Frames without calibration and period information. Old Data Format Frames are 2020 bytes long (the "28-byte" sections are 100 20-byte sections [Table 4] because bytes 5-12 are not present).

TABLE 4: DATA FORMAT FRAME (CONTINUED)

Field	Format	Description
number of channels	4-byte integer	number of channels in Data Frame
frame time length	4-byte integer	time in milliseconds that each Data Frame following will represent
site and channel flags, calibrations, and names	100 28-byte[1] sections:	
	4-byte integer	flags byte 1: authentication (0 = off; 1 = on) byte 2: compression (0 = no compression; 1 = Canadian compression applied after authentication; other values as required) bytes 3 and 4: unused
	4-byte IEEE float[2]	bytes 5–8 calibration factor (nm/count)
	4-byte IEEE float	bytes 9–12 calibration period (seconds)
	16 bytes ASCII	bytes 13–28: site/channel names with slash character required as separator; string begins at first byte and contains no spaces; remaining bytes zeroed; unused strings zeroed; site/channel designations are required to appear in the order the data will appear in the Data Frames

1. All 28 bytes are zero for unused channels. No channels may be specified after the first unused channel. Unused channels will not appear in the Data Frames that follow a Data Format Frame.
2. The units for calibration of seismic sensors are nm/count. The units for other sensors are defined as required. For example, infrasonic sensors use Pa/count.

DATA FRAME

The Data Frame contains the raw time series data. In normal operation, Data Frames comprise the bulk of the transmission. Each Data Frame consists of a single Data Frame Header and multiple Channel Subframes. Figure 2 illustrates this relationship. The Data Frame Header contains size and time information and possibly weather or other state of health data. The Channel Subframes contain the actual data along with an authentication signature if appropriate. The number and order of Channel Subframes match the number and order specified in the Data Format Frame.

.

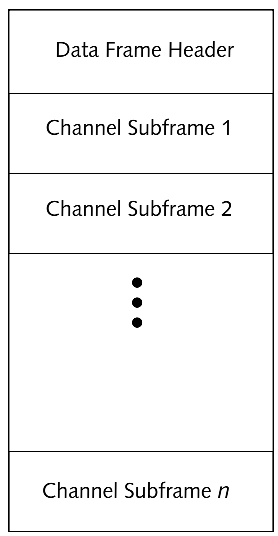


FIGURE 2. DATA FRAME COMPONENTS

Each Data Frame is transmitted only once. If it becomes corrupted, it is rejected by the dc and not replaced. Data Frames must be short enough to allow near-real-time transmission and to limit data losses due to errors. They must also be long enough to be processed efficiently. The size of the Data Frame is chosen by the dp, but must not exceed 10,000,000 bytes. Data Frames typically contain 10 to 30 seconds of data for each channel. Shorter Data Frames are sometimes used to send incomplete data bordering a data outage. Shorter and longer Data Frames have been used at some dps due to requirements of the station system.

Data Frame Header

The Data Frame Header supplies information that is common to each channel. All portions of this format that are of arbitrary length are required to be divisible by four; this rule saves unnecessary byte manipulation for machines with 4-byte word alignment. Table 5 defines the Data Frame Header.

TABLE 5: DATA FRAME HEADER

Field	Format	Description
frame identification	4-byte integer	identification for Data Frame = 1
frame length	4-byte integer	length in bytes of Data Frame, including header and all Channel Subframes
nominal time	8-byte IEEE float	nominal start time of all channels in frame, in seconds since January 1, 1970, 00:00, and not counting leap seconds
description length	4-byte integer	number of bytes in the following description (maximum 1,024, divisible by 4, 0 is default)
description[1]	description-length bytes	description (for example, maintenance or state-of-health)

1. Any sequence of bytes is permitted in the description field, so arbitrary comments and state-of-health information are possible. However, dc processing for this field is undefined. Dcs are unlikely to have any special handling for this field. Use of this field for time series such as weather data is explicitly discouraged; the preference is to use a data channel.

Channel Subframe

The Channel Subframes portion of the Data Frame provide the raw data for each channel. The set of Channel Subframes follow the Data Frame Header and are arranged in the order specified in the Data Format Frame. Table 6 defines the Channel Subframe.

To avoid unnecessary communication overhead, the authentication field is collapsed to zero bytes for channels that are not flagged for authentication in the Data Format Frame.

TABLE 6: CHANNEL SUBFRAME

Field	Format	Description
packet length	4-byte integer	length of packet in bytes, not counting this word, for channel that follows (divisible by 4)
authentication	40-byte string (or 0 bytes)	authentication signature for this Channel Subframe; zero length field if not enabled
time stamp	8-byte IEEE float	seconds since January 1, 1970, 00:00, (excluding leap seconds) for first sample; frame time must be within the time length (specified in the Data Format Frame) of the nominal time (specified in the Data Frame Header)
samples	4-byte integer	number of samples in channel packet

TABLE 6: CHANNEL SUBFRAME (CONTINUED)

Field	Format	Description
status	4-byte string	data status byte (most significant byte): bit 31 1 = dead channel bit 30 1 = zeroed data bit 29 1 = clipped bit 28 1 = calibration signal bit 24–2 undefined station status byte: bit 23 1 = vault door open bit 22 1 = authentication box opened bit 21 1 = equipment moved bit 20 1 = clock differential too large bit 16–19 undefined station specific bits: bit 0–15 user defined (for example, station status counter)
data	4-byte integers or Canadian compression scheme	data for channel

The 32-bit status word contains indicators of the status of the station equipment or data. To allow the dc to monitor the completeness of each channel data stream, all data should be sent. This can happen if they are zeroed or clipped or if calibration signals are present. The dc is responsible for handling these conditions properly.

ALERT FRAME

Alert Frames are sent to the dc or dp to notify them of a change in the current flow of data. An Alert Frame precedes any change that can be anticipated, including the dp announcing a termination to the dc before disconnecting, the dp changing the data format, the dc announcing an immediate termination, or the dc changing the address and port of the connection. More alert types may be added as this standard evolves. Table 7 defines the Alert Frame.

When the dp changes the data format (alert type 2), the Alert Frame must be followed by a Data Format Frame and then the new Data Frames. When the dc is alerting the dp of a change in address and port (alert type 1001), the Alert Frame must be followed by a Port Assignment Frame.

TABLE 7: ALERT FRAME

Field	Format	Description
frame identification	4-byte integer	identification for Alert Frame = 4
frame length	4-byte integer	length in bytes of Alert Frame
alert type	4-byte integer	key for alert type 0 = no action 1 = immediate termination by dp 2 = flush format, wait for new Data Format Frame (from dp) 1000 = immediate termination by dc 1001 = new connection information (from dc)

EXAMPLE

This section describes a hypothetical continuous data session between an IMS station (dp) and the IDC (dc). The dp is assumed to be a station whose computer has recently returned to operation after a power outage and is seeking a connection to the IDC. The hypothetical station name is ZZA0, a nine-element seismic array having nine short-period vertical seismometer channels sampled at 40 Hz and a three-component (3-C) broadband seismometer at element ZZA0 sampled at 80 Hz.

The software managing the sending of the station data first attempts to open a connection to `ts.idc.org` on port `8000`. The time-out is set to 90 seconds. If the time-out expires because communications are not fully established after restarting the station systems, the program retries the connection. When successful, a TCP/IP connection is established. The dp immediately sends the Connection Request Frame as shown in Figure 3, where "0" is an ASCII alphanumeric 0, and (nil) is a zero byte.

1	2	3	4	5	6	7	8
Z	Z	A	0	(nil)	(nil)	(nil)	(nil)

FIGURE 3. EXAMPLE CONNECTION REQUEST FRAME

While waiting for the dc to respond, the dp again sets a 90-second time-out. This time-out may be exceeded for the following reasons:

- communications may have failed
- the dc may reject the station name or the originating address
- the dc may be slow and fail to meet the deadline

If the time-out is exceeded, the dp closes the connection and attempts to re-connect and send a new Connection Request Frame. Normally, the dc sends a Port Assignment Frame to the dp within 90 seconds (see Figure 4).

1	2	3	4	5	6	7	8
1	4	0	.	1	6	2	.
1	.	5	/	7	1	4	4
(nil)	(nil)	(nil)	(nil)	(nil)	(nil)	(nil)	(nil)
(nil)	(nil)	(nil)	(nil)	(nil)	(nil)	(nil)	(nil)

FIGURE 4. EXAMPLE PORT ASSIGNMENT FRAME

The Port Assignment Frame shown in Figure 4 tells the dp to connect to the Internet address `140.162.1.5` on port `7144`. After receiving the Port Assignment Frame, the dp immediately drops the original connection to the dc, tries connecting to the new address and port, and sets a time-out of 600 seconds for this connection to be established. If the connection fails, the dp starts the process again by connecting to `ts.idc.org` on port `8000` and sending a Connection Request Frame.

After connecting to the dc on the assigned port, the dp sends a Data Format Frame. Figure 5 shows an example Data Format Frame with integers and floats represented in hexadecimal and strings represented in ASCII. The channel order reflects the order in which the data will be transmitted (short-period first, followed by broadband). The first 20 bytes give the frame identification (2 for Data Format Frame), frame length (2,820 bytes), maximum Data Frame size (19,440 bytes), number of channels (12), and the number of milliseconds per Data Frame (10,000).

The maximum Data Frame size is computed by assuming at most three bytes per sample, three 80-Hz, and nine 40-Hz channels (with 100 bytes additional information per channel), 20 bytes required header, plus 180 bytes of descriptive data (this dp will limit itself to 180 bytes in that field). The ZZA0/SHZ data have authentication signatures included, and all data are compressed using the Canadian compression scheme.

1	2	3	4	5	6	7	8
frame identification (2)				**frame length (2,820)**			
00	00	00	02	00	00	0B	04
maximum frame size (19,440)				**number of channels (12)**			
00	00	4B	F0	00	00	00	0C
frame time length (10,000)				**channel 1 flags**			
00	00	27	10	01	01	00	00
channel 1 calibration factor (0.196)				**channel 1 calibration period (1.0)**			
3E	48	B4	39	3F	80	00	00
channel 1 station and channel designation							
Z	Z	A	0	/	S	H	Z
(nil)	(nil)	(nil)	(nil)	(nil)	(nil)	(nil)	(nil)
channel 2 flags				**channel 2 calibration factor (0.196)**			
00	01	00	00	3E	48	B4	39
channel 2 calibration period (1.0)				**channel 2 station and channel designation**			
3F	80	00	00	Z	Z	A	1
(remainder of channel 2 and channel 3–11 information)							
channel 12 station and channel designation							
Z	Z	A	0	/	B	H	Z
(nil)	(nil)	(nil)	(nil)	(nil)	(nil)	(nil)	(nil)
(channel 13–100 all bytes nil)							

FIGURE 5. EXAMPLE DATA FORMAT FRAME

After the Data Format Frame is sent, the dp can send Data Frames to the dc. Figure 6 shows the first part of an example Data Frame.

1	2	3	4	5	6	7	8
frame identification (1)				**frame length (7,500)**			
00	00	00	01	00	00	1D	4C
nominal time (891043200.000)							
41	CA	8E	1E	C0	00	00	00
description length (0), end of header				**length of first Channel Subframe (536)**			
00	00	00	00	00	00	02	18

FIGURE 6. EXAMPLE DATA FRAME HEADER

The Data Frame Header includes the frame identification (1), the frame length (7,500), and the nominal time (an 8-byte float). The data frame length in the example is less than the maximum data frame length specified in the Data Format Frame. This is not unusual, especially with compressed data. In this example, compression was very effective, resulting in an average of 1.2 bytes per sample. This Data Frame also contains nothing in the descriptive field, which is also not unusual. The last 4 bytes shown are not part of the header; they are the first field of the first Channel Subframe (packet length). In this case, the 400 samples of ZZA0/SHZ data are compressed to 480 bytes. With the authentication signature (this is the only signed channel in the example), and the remaining 16 subframe header bytes, the length is 536 bytes. The packet length would be followed by the authentication signature, the actual time of the first sample, the number of samples (40), the status, and the raw data bytes. The remaining Channel Subframes would follow in the order specified in the Data Format Frame.

In this scenario, the data are sent continuously from the dp to the dc until a time when maintenance is scheduled for the broadband seismometers. The communications are left running during this maintenance procedure by instructing the sending software to interrupt communications and send a new format. First, the dp sends an Alert Frame to the dc (Figure 7).

1	2	3	4	5	6	7	8
00	00	00	04	00	00	00	0C
00	00	00	02				

FIGURE 7. EXAMPLE ALERT FRAME

The Alert Frame notifies the dc to pause and expect a new format. The dp then sends a new Data Format Frame with 9 channels specified and resumes sending data (this time without the broadband channels). When the scheduled maintenance of the broadband channels is completed, the dp again sends an Alert Frame to the dc followed by a new Data Format Frame. The transmission of Data Frames then commences, with the 3 broadband channels returned.

Chapter 3: Data Formats

This chapter describes the data formats for continuous data exchange and includes the following topics:

- Introduction
- Uncompressed Format
- Canadian Compressed Format

Chapter 3: Data Formats

INTRODUCTION

Continuous time series data may be transmitted in uncompressed or Canadian compressed formats and may contain authentication signature information. All data formats transfer data without loss of information. The Canadian compression scheme reduces the cost of data transmission, however, because fewer bits are required to transmit the same number of uncompressed data samples.

UNCOMPRESSED FORMAT

The uncompressed format specifies 4-byte integer representation. Use of IEEE representation elsewhere implies forward byte ordering of all multi-byte values. Values generated on machines that do not use IEEE or forward byte ordering as their native numerics may need to swap bytes before transmitting data.

CANADIAN COMPRESSED FORMAT

The Canadian compression scheme is often used for continuous data transmission. Like many "lossless" compression schemes, it is based on an indexing of the second difference of the data.

Difference Scheme

Let "S" be the sampled data so that S(1) is the first sample, S(2) is the second sample, and so on. The samples, S, are represented as 32-bit, 2's-complement integers, with a range of -2^{31} to $+2^{31}-1$. Any data sample whose value is unknown is given an integer value of -2^{31}, or 0 if data are compressed.

Define "D" as the first difference between two subsequent samples so that:

$$D(2) = S(2) - S(1)$$
$$D(3) = S(3) - S(2)$$
$$\ldots$$
$$D(j) = S(j) - S(j-1)$$

Further define "D_2" as the second difference between two subsequent first difference samples so that:

$$\begin{aligned} D_2 &= D(j) - D(j-1) \\ &= S(j) - 2S(j-1) + S(j-2) \end{aligned}$$

A sequence of N+1 numbers consisting of S, D, and D_2 values constitutes a packet of data representing N data values:

$$S(1), D(2), D_2(3), D_2(4), \ldots, D_2(N+1)$$

or, in terms of S:

$$S(1), S(2) - S(1), S(3) - 2S(2) + S(1), \ldots, S(N+1) - 2S(N) + S(N-1)$$

N+1 samples are in the data block. The last sample in the sequence contains the value of the first sample in the next data packet and is used as an independent error check of the compression and decompression processes.

Index Scheme

The "N" difference values of the sequence derived above (the first sample is not a difference) are represented as 4, 6, 8, 10, 12, 14, 16, 18, 20, 24, 28, or 32-bit, 2's-complement integers with ranges from –8 to +7 for four bits, –32 to +31 for six bits, and so on, up to -2^{31} to $+2^{31}-1$ for 32 bits.

An index is used to specify how many bits are used to encode the difference values of the sequence. The index contains a number of 16-bit index entries, one for each block of 20 differences. Thus, a 10-second data packet sampled at 40 samples per second will have an index containing 20 entries totaling 40 bytes.

Each index entry encodes the bit lengths of the difference values of the sequence in groups of four as in Table 8.

TABLE 8: 16-BIT INDEX

Most Significant Bit															Least Significant Bit
h	a	a	a	b	b	b	c	c	c	d	d	d	e	e	e

- h = 0

  all 20 samples can be encoded in ≤ 18 bits per sample
- h = 1

  at least one sample needs ≥ 20 bits for encoding
- aaa

  encodes the number of bits per sample required to specify the first four data sample differences (for example, for samples 2 through 5); its interpretation depends on whether h = 0 or 1 (see Table 9)
- bbb

  bits per sample encoding for samples 6 through 9
- ccc

  bits per sample encoding for samples 10 through 13
- ddd

  bits per sample encoding for samples 14 through 17
- eee

  bits per sample encoding for samples 18 through 21

Table 9 shows the number of bits assigned to samples based on the three-bit sample codes and the value of h.

TABLE 9: BIT LENGTHS FOR VARIOUS SAMPLE CODES

3-bit Sample Code	Sample Bit Length	
	h = 0	h = 1
000	4	4
001	6	8
010	8	12
011	10	16
100	12	20
101	14	24
110	16	28
111	18	32

Chapter 4: Policies for Continuous Data Transmission

This chapter describes the policies for continuous data transmission and includes the following topics:

- Field Values
- Descriptive Information
- Size and Time Limits
- Grouping and Ordering
- Other Policies

Chapter 4: Policies for Continuous Data Transmission

FIELD VALUES

The valid values for key fields of the protocol are indicated in Table 10.

TABLE 10: VALID VALUES FOR PROTOCOL FIELDS

Fields	Table	Type	Default	Values	Example
station name	2, 4	ASCII	required	ASCII alpha-numeric string, left justified, padded with ASCII nulls; first character must be alphabetic	`WRA`
IP address	3	ASCII	required	ASCII alpha-numeric string, left justified, padded with ASCII nulls, of form nnn.nnn.nnn.nnn	`140.162.2.90`
port	3	integer	required	0-65535	`13282`
nominal time, time stamp	5, 6	IEEE double	required	time in seconds since 1 January 1970 00:00:0.00	`990001234.125`
site name	4	ASCII	required	ASCII alpha-numeric string, left justified, padded with ASCII nulls; first character must be alphabetic	`WB2`
channel name	4	ASCII	required	FDSN channel name	`BHZ`

DESCRIPTIVE INFORMATION

These policies pertain to ancillary information not defined by the protocol but required by application software.

Static Configuration Data

Descriptive information is required to configure software to receive and process continuous data. These data will be provided separately from this protocol. The exact list is not specified, but will include the IP address of the sending computer, station codes, instrument frequency response, and so on.

Unknown Station, Site, and Channel Names

Data that are received with station, site, or channel codes that have not been included in the static configuration database are discarded.

SIZE AND TIME LIMITS

These policies pertain to constraints on Data Frames.

Maximum Number of Channels

Up to 100 channels can be transmitted as part of a single Data Frame.

Time Synchronization

All Channel Subframes of a Data Frame will have the same start time and duration to within one sample time unit. There is one Channel Subframe for each channel in a Data Frame.

Duration and Size of Channel Subframe

The IMS requires that Data Frames contain ≤ 10 seconds of data for seismic (short-period and broad-band) and hydroacoustic sensors and ≤ 30 seconds for infrasonic and seismic long-period data. However, to avoid excessive authentication processing load and frame processing overhead, it is recommended that frames contain exactly 10 seconds or 20 seconds of data, respectively. Because each channel in a Data Frame has one Channel Subframe, these constraints apply to the Channel Subframe as well.

For illustrative purposes, it is interesting to see the sizes of the various elements of a Channel Subframe. Table 11 shows the approximate sizes, in bytes, for the stated conditions, using 40 sps, 240 sps, and 10 sps for the respective instrument types, and assuming a 3-byte digitizer and 50 percent reduction from data compression.

TABLE 11: REPRESENTATIVE SIZES OF CHANNEL SUBFRAMES

Fields	Seismic	Hydroacoustic	Infrasonic
data	600	3,600	450
signature	40	40	40
description and status	20	20	20
subframe total	660	3,660	510

Applications should be able to handle 1000-second Channel Subframes.

Duration and Size of Data Frame

The duration of the Data Frame will be the same as that of the Channel Subframes, described in the previous section. The total size of the Data Frame will be the sum of the sizes of the Channel Subframes plus approximately 20 bytes for the frame header and trailer. In the case of a four-element infrasonic array, the total size is approximately 2,060 bytes. In the case of a 10-element seismic array, the total size is approximately 6,620 bytes.

The IMS requirements on frame duration hold for continuous transmission of real time data from the seismic, hydroacoustic, and infrasonic instruments. This condition does not apply in three situations: (1) the transmission of old data, which may accumulate during an outage; (2) the transmission of ancillary data, such as weather readings; and (3) transmission of data segments, if policy permits such use of the protocol. Policy has not been established for these cases.

GROUPING AND ORDERING

This section describes policies related to the grouping of Channel Subframes into Data Frames and ordering of transmission of frames.

Mixed Sample-rate Data

Channel Subframes from a single station with different sample rates shall be included in a single Data Frame, as long as all other policies are satisfied. For seismic arrays, this policy requires inclusion of broad-band and short-period data in the same Data Frame.

Data Ordering

During normal operations, data are sent from the provider to the consumer as soon as they become available. When service has been interrupted, the policy is to send the most recent data first and to transmit older data, if any accumulated during the outage, as bandwidth permits. When older data are transmitted in reverse time order the data ordering scheme is referred to as Last In, First Out (LIFO). Because dcs find fragmentary data to be unusable for processing and analysis, a modified LIFO scheme is followed. During periods of slow communications, data continue to be sent as time-ordered frames, falling behind the most recent frames, which are stored in a buffer. When a pre-determined number of frames have built up in the buffer, the transmission process jumps to the most recent (or last in) frame. To recover, the software sends several neighboring backlog frames in a series before returning to serve the recent data. These actions create the concept of a "super frame", the minimum number of contiguous frames that represent useful data for processing at a dc (typically about 10 minutes total length).

TCP/IP guarantees that frames will be received in the same order they are sent.

Use of Data Format Frames

The design of this protocol requires some discussion of the use of Data Format Frames. Data Format Frames in CD-1 are designed to separate configuration and indexing information from the raw data. The purpose for this is both to reduce transmission of redundant information and to reduce the overhead in the processing of such information at the dc. Due to this design, it is often computationally expensive to process Data Format Frames at the dc. Therefore, transmission of Data Format Frames should be limited to the minimum required.

To reduce Data Format Frames, the following actions should be taken:

- Do not transmit redundant Data Format Frames. If the Data Format Frame is the same as the last, drop it, and just continue sending Data Frames.
- Always send channels in the same order.
- Unless required by a dead channel or a very long intra-site communications failure at the dp, always send the same number of samples.
- Send all channels together in a single Data Frame.
- If sending LIFO data from a backlog, send several backlog frames immediately following each other, if possible. This is related to the concept of "modified LIFO".
- The CD-1 protocol is only "near" real time. Consider a short delay of several seconds to allow a complete Data Frame to be assembled and sent.

- Do not mix arrays in a single stream. For example, do not mix seismic and infrasonic Channel Subframes as single Data Frames. The CD-1 concept is built around a single sensor array in a single data stream. Interleaving format and data frame sequences generated for separate arrays is incompatible with the protocol.

OTHER POLICIES

This section describes miscellaneous policies.

Refusal of Connection Request

If the dc denies a connection request from a dp, no response is returned to the requester.

Station and Site Names

The station name identifies the location of a sensor site or a collection of sensor sites. The station name for seismic arrays is the beam point for signal processing. The site name is the location of a specific sensor.

Seismic station and site names shall be registered with the International Seismic Centre (ISC) or the National Earthquake Information Center (NEIC) of the U.S. Geological Survey. Seismic station and site names shall conform to standards given in [USG85]. For example, station names shall be three to five characters in length, and therefore the station field will never be completely filled. No counterpart registry exists for hydroacoustic and infrasonic station or site names, but station and site name formats shall conform to IMS station naming conventions [IMS01].

Channel Names

The channel name is the three-character string that labels the bandwidth, gain, and orientation of the sensor. Channel names for seismic stations shall conform to the standards of the Federation of Digital Seismograph Networks (FDSN) [IRI93]. Channel names that are not defined by the FDSN shall conform to IMS channel naming conventions [IMS01].

Disconnection Due to Time-out

A time-out is defined to be an elapsed interval of time when no Frames have been sent. The time-out is defined in Table 1 on page 3. Under these conditions both the dp and dc will disconnect. This policy provides a well-defined failure mode in the event of an unrecoverable fault in either the hardware or application software at either end of the link.

Duplicate and Overlapping Data

Data Frames containing data that duplicate or overlap any data successfully received previously are discarded on a channel-by-channel basis without notifying the dp.

Authentication of Data

Compression of data is to be applied after authentication signatures are calculated. The uncompressed data format used for signature calculation must be the data type specified in the Channel Subframe.

References

The following sources supplement or are referenced in document:

[Gan79] Gane, C., and Sarson, T., *Structured Systems Analysis: Tools and Techniques*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1979.

[GSE95b] Group of Scientific Experts, *GSETT 3 Documentation, Volume Two: Operations,* CRP/243, 1995.

[IDC3.4.3Rev0.2] Science Applications International Corporation, *Formats and Protocols for Continuous Data CD-1.1*, *Revision 0.2,* SAIC-01/3027Rev0.2, 2001.

[IDC7.4.1] Science Applications International Corporation, *Continuous Data Subsystem CD–1.1*, SAIC-01/3012, 2001.

[IMS01] International Monitoring System Division, Operations and Maintenance Group, IMS-OM-2001-05, 2001.

[IRI93] Incorporated Research Institutes for Seismology, *Standards for the Exchange of Earthquake Data (SEED)*, Reference Manual SEED Format Version 2.3, February, 1993
online: `http://www.iris.washington.edu/manuals/acrobat/T_SEED.pdf`

[USG85] U.S. Geological Survey, *Seismograph Codes and Coordinates*, Open File Report 85-719, 1984
online: `http://wwwneic.cr.usgs.gov/neis/station_book/background.html`

Glossary

Symbols

3-C

Three-component.

A

array

Collection of sensors distributed over a finite area (usually in a cross, triangle, or concentric pattern) and referred to as a single station.

ASCII

American Standard Code for Information Interchange. Standard, unformatted 256-character set of letters and numbers.

authentication signature

Series of bytes that are unique to a set of data and that are used to verify the authenticity of the data.

B

byte alignment

Alignment of stored values in memory so that they begin at byte boundaries evenly divisible by some number, commonly 4.

C

calibration

Process of determining the response function and sensitivity of an instrument or its derived channel.

Canadian compression

Lossless time-series compression method for 4-byte integers that was originally developed for the Canadian National Seismic Network and that allows a fixed number of samples per block.

channel

Component of motion or distinct stream of data.

client

Software module that gathers and presents data to an application; it generates requests for services and receives replies. This term can also be used to indicate the requesting role that a software module assumes by either a client or server process.

connection

Open communication path between protocol peers.

continuous waveform data

Waveform data that are transmitted to the IDC on a nominally continuous basis.

count(s)

Units of digital waveform data.

D

data consumer

Receiver of CD-1 or CD-1.1 Data Frames. This is always a data center.

data provider

Sender of CD-1 or CD-1.1 Data Frames. This may be a station or a data center that forwards data.

E

element

(1) Single station or substation of a sensor array, referred to by its element name (such as YKR8), as opposed to its array name (YKA in this example). (2) Data storage location in a data array.

epoch time

Number of seconds after January 1, 1970 00:00:00.0.

F

FDSN

Federation of Digital Seismic Networks.

frame

Logical collection of digital information that is transmitted as a unit from application to application.

G

GSETT

Group of Scientific Experts, Technical Test. The GSE addressed technical issues relating to test-ban negotiations. The technical test was a full-scale test of a system to monitor the treaty.

H

hydroacoustic

Pertaining to sound in the ocean.

Hz

Hertz.

I

IDC

International Data Centre.

IEEE

Institute for Electrical and Electronic Engineers.

IMS

International Monitoring System.

infrasonic (infrasound)

Pertaining to low-frequency (sub-audible) sound in the atmosphere.

IP

Internet protocol.

IP address

Internet Protocol address, for example: 140.162.1.27.

ISC

International Seismological Centre.

L

LIFO

Last in, first out. A frame-ordering scheme for sending data built up in a buffer during communications outages. Recent frames are sent before older frames.

N

NEIC

National Earthquake Information Center.

nil

Empty set.

nm

Nanometers.

P

Pa

Pascals.

port

Connection to a computer.

S

sample

Any physical entity counted on a detector.

seismic

Pertaining to elastic waves traveling through the earth.

sensor

Device that detects a variable physical quantity and converts it to an electrical signal that can be digitally sampled.

server

Software module that accepts requests from clients and other servers and returns replies.

site

Location of a sensor within a station.

station

Collection of one or more monitoring instruments. Stations can have either one sensor location (for example, BGCA) or a spatially distributed array of sensors (for example, ASAR).

T

TCP/IP

Transmission Control Protocol/Internet Protocol.

time, epoch

See epoch time.

time series

Time ordered sequence of data samples. Typically a waveform or derived from waveforms, such as a beam.

U

UNIX

A portable, multi-user, time-shared operating system that supports process scheduling, job control, and a programmable user interface.

Index

A

C

D

E

F

I

K

L

M

N

P

R

S

T

U